

CLAIMS

What is claimed is:

1. A database of unique nucleotide sequences, said database comprising nucleotide
sequences greater than about 100 nucleotides in length.
2. A database of unique nucleotide sequences, said database comprising nucleotide
sequences between about 100-500 nucleotides in length.
3. A database of unique nucleotide sequences, said database comprising nucleotide
sequences between about 100-1000 nucleotides in length.
4. The database of any of claims 1-3, wherein said nucleotide sequence is a
deoxyribonucleotide sequence.
5. The database of any of claims 1-3, wherein said nucleotide sequence is a ribonucleotide
sequence.
6. The database of any of claims 1-3, wherein said sequences are derived from animal DNA
or RNA.
7. The database of claim 6, wherein said animal is a human.
8. The database of claim 6, wherein said animal is a mouse.

9. The database of any of claims 1-3, wherein said sequences are derived from plant DNA or RNA.

10. The database of any of claims 1-3, wherein said plant is a single-cell plant.

11. The database of any of claims 1-3, wherein said sequences are derived from fungal DNA or RNA.

12. The database of any of claims 1-3, wherein said sequences are derived from DNA or RNA of a microorganism or virus.

13. The database of any of claims 1-3, wherein said sequences are derived from DNA or RNA of a single-cell eukaryote.

14. The database of any of claims 1-3, wherein said sequences are derived from synthetic man-made DNA or RNA.

15. The database of any of claims 1-3, wherein said sequences are postulated based upon amino acid sequences.

16. The database of any of claims 1-3, wherein said database is encoded in a biological medium.

17. The database of any of claims 1-3, wherein said database is encoded in a written medium.

18. The database of any of claims 1-3, wherein said database is encoded in an electronic medium.

19. The database of claim 18, wherein said electronic medium is a computer-readable medium.

20. The database of claim 19, wherein said computer-readable medium is addressable through an internet connection.

21. A kit for analyzing nucleotide sequences comprising:

an electronic medium readable by a computer, said medium encoding a database of unique nucleotide sequences, said database comprising nucleotide sequences greater than about 100 nucleotides in length.

22. A kit for analyzing nucleotide sequences comprising:

an electronic medium readable by a computer, said medium encoding a database of unique nucleotide sequences, said database comprising nucleotide sequences greater than about 100 nucleotides in length; and,

instructions for the use of said database.

23. A kit for analyzing nucleotide sequences comprising:

an electronic medium readable by a computer, said medium encoding a database of unique nucleotide sequences, said database comprising nucleotide sequences greater than about 100 nucleotides in length;

instructions for the use of said database; and,

a computer.

24. An improved database of nucleotide sequences, said database comprising nucleotide sequences greater than about 100 nucleotides in length, wherein said improvement consists entirely of only unique nucleotide sequences entered into said database only one time.

25. A computer-generated database consisting of only unique nucleotide sequences, said database comprising nucleotide sequences greater than about 100 nucleotides in length.

26. A method for generating a database of sequences that are greater than or equal to about 100 nucleotides in length, wherein each sequence is entered into the database only one time, the method comprising the steps of :

selecting a query sequence from a redundant database;

masking said query sequence with known repeat sequences;

comparing said masked query sequence with identified unique sequences;

identifying a unique portion of the query sequence that does not have a similar sequence in any of the identified unique sequences; and

adding the unique portion of the query sequence to a unique database.

27. A database product of the process of claim 26.

28. The method of claim 26, wherein said sequence is a deoxyribonucleotide sequence.

29. The method of claim 26, wherein said sequence is a ribonucleotide sequence.

30. The method of claim 26, wherein said sequences are derived from animal DNA or RNA.

5

31. The method of claim 30, wherein said animal is a human.

32. The method of claim 30, wherein said animal is a mouse.

10 33. The method of claim 26, wherein said sequences are derived from plant DNA or RNA.

34. The method of claim 33, wherein said plant is a single-cell plant.

35. The method of claim 26, wherein said sequences are derived from fungal DNA or RNA.

15

36. The method of claim 26, wherein said sequences are derived from DNA or RNA of a microorganism or virus.

37. The method of claim 26, wherein said sequences are derived from DNA or RNA of a single-cell eukaryote.

20

38. The method of claim 26, wherein said sequences are derived from synthetic man-made DNA or RNA.

39. The method of claim 26, wherein said sequences are postulated based upon amino acid sequences.

40. The method of claim 26, wherein said database is encoded in a biological medium.

41. The method of claim 26, wherein said database is encoded in a written medium.

42. The method of claim 26, wherein said database is encoded in an electronic medium.

43. The method of claim 42, wherein said electronic medium is a computer-readable medium.

44. The method of claim 43, wherein said computer-readable medium is addressable through an internet connection.

45. The method of claim 26, wherein said redundant database is a Public Domain Database.

46. The method of claim 45, wherein said Public Domain Database is GenBank.

47. The method of claim 45, wherein said Public Domain Database is dbEST.

48. The method of claim 45, wherein said Public Domain Database is TIGR.

49. The method of claim 45, wherein said Public Domain Database is SwissProt.

50. The method of claim 26, wherein said comparing step further utilizes a Database Search Algorithm.

51. The method of claim 50, wherein said Database Search Algorithm is BLAST.

52. The method of claim 50, wherein said Database Search Algorithm is FASTA.

53. The method of claim 50, wherein said Database Search Algorithm is Smith-Waterman.

54. The method of claim 26, wherein said comparing step further utilizes a Scoring Matrix Program.

55. The method of claim 54, wherein said Scoring Matrix Program is PAM.

56. The method of claim 54, wherein said Scoring Matrix Program is BLOSUM.

57. The process of Figure 1A.

58. The process of Figure 1B.

59. The process of Figure 2.

60. A method for identifying unique nucleotide sequences, the method comprising the steps of:

selecting a query sequence from a redundant database file;

comparing the query sequence with a repeat database file and a unique database file;

analyzing the results of the comparison of the query sequence with the repeat database file and the unique database file to determine if there is one or more nucleotide sequences within the repeat database file and the unique database file that match a nucleotide sequence within the

5 query sequence; and

identifying any unique nucleotide sequences within the query sequences that do not match any nucleotide sequence within the repeat database file and the unique database file.

09033056 082001